EXHIBIT E

AI and Veterinary Medicine: Performance of Large Language Models on the North American Licensing Examination

Mirana Angel
Institute for Geomics and Bioinformatics
University of California Irvine
Irvine, USA
mcangel@uci.edu

Anuj Patel
Department of Computer Science
University of California Irvine
Irvine, USA
patelad2@uci.edu

Haiyi Xing
Department of Computer Science
University of California Irvine
Irvine, USA
haiyix2@uci.edu

Dylan Balsz
Internal Medicine
Anivive Life Sciences
Long Beach, USA
dylan@anivive.com

Cody Arbuckle
Internal Medicine
Anivive Life Sciences
Long Beach, USA
cody@anivive.com

David Bruyette
Internal Medicine
Anivive Life Sciences
Long Beach, USA
david@anivive.com

Pierre Baldi
Department of Computer Science
University of California Irvine
Irvine, USA
pfbaldi@uci.edu

2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) | 979-8-3503-1890-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/SNAMS60348.2023.10375414

***Abstract*—This study aimed to assess the performance of Large Language Models on the North American Veterinary Licensing Examination (NAVLE) and to analyze the impact of artificial intelligence in the domain of animal healthcare. For this study, a 200-question NAVLE self-assessment sourced from ICVA's website was used to evaluate the performance of three language models: GPT-3, GPT-4, and Bard. Questions involving images were omitted leaving a 164 text-only sample exam. Results were analyzed by comparing generated responses to the answer key, and scores were assigned to evaluate the models' veterinary medical reasoning capabilities. Our results showed that GPT-4 outperformed GPT-3 and Bard, passing the exam with 89 % of the text-only questions correctly. GPT-3 and Bard only achieved an accuracy of 63.4 % and 61 % respectively on the same set of questions. Language models hold promise for enhancing veterinary practices through expanded educational opportunities in the veterinary curriculum, improved diagnostic accuracy, treatment times, and efficiency. However, potential negatives include challenges in changing the current educational paradigm, reduced demand for professionals or paraprofessional concerns surrounding machine-generated decisions. Responsible and ethical integration of language models is crucial in veterinary medicine.**

***Index Terms*—Artificial Intelligence, LLM, ChatGPT, Bard, Veterinary Medicine, Medical Education, Societal Impact**

I. Introduction

In recent years, the rapid growth of artificial intelligence (AI) has significantly influenced various industries, including healthcare. The development of increasingly powerful AI models, such as large language models (LLMs) has facilitated the automation of diverse tasks and the enhancement of decision-making processes. Consequently, the adoption of AI technology has emerged as a pivotal factor in gaining a competitive edge and boosting efficiency across industries [1]. Here we provide an initial assessment of the applicability of LLMs in veterinary medicine by testing their ability to pass a standard veterinary education test.

The veterinary field encompasses a wide array of professions and specializations, all dedicated to the care and well-being of animals. Veterinarians, who are extensively trained to diagnose and treat various conditions in numerous species ranging from domesticated animals and livestock to wildlife, are a cornerstone of this field. As the veterinary field continues to evolve, new technologies and techniques are revolutionizing the diagnosis and treatment of animal health issues [2].

The advent of diverse AI technologies, such as state-of-the-art text, sound, image, and video data analysis algorithms, have significantly advanced veterinary medicine in areas such as disease diagnosis, treatment planning, and precision medicine [2, 3, 4]. However, current AI models are typically task-specific and lack the capability for independent medical reasoning [5]. This limitation has prompted researchers to explore the potential of large language models, which have demonstrated remarkable cognitive reasoning abilities, in addressing these shortcomings in all fields.

Among large language models, Generative Pre-trained Transformer (GPT) and Bard have emerged as frontrunners, exhibiting outstanding performance in various applications [6, 7, 8]. GPT-3 and GPT-4, as well as Bard, adopt the decoder-only architecture of the transformer model [9]. GPT-3 encompasses 175 billion parameters and showcases remarkable versatility across a range of tasks. In an advancement over GPT-3, GPT-4 boasts an unprecedented one trillion parameters, addressing many of the limitations previously associated with GPT-3. Both GPT iterations were pre-trained on extensive text corpora and subsequently fine-tuned for specialized tasks [6, 7].

Concurrently, Google's Bard initially employed the Language Model for Dialogue Applications (LaMDA), with 137 billion parameters, primarily pre-trained on public dialog data and web text [8]. However, in a recent update on May 11, 2023, Google made a transition to the PaLM-2 model for Bard, elevating its parameter count to 540 billion. This enhancement is anticipated to imbue Bard with improved conversational comprehension and response accuracy, particularly bolstering its capabilities in medical reasoning [10].

Impressively, GPT-3 has successfully passed the United States Medical Licensing Examination (USMLE) [11]. The USMLE is a set of three standardized tests of expert-level knowledge, which are required for medical licensure (MD) in the United States. These authors found that ChatGPT performed at or near the passing threshold of 60% accuracy. Even more recently, we were able to demonstrate that GPT-4 was able to pass the specialized and challenging American Board of Anesthesiology (ABA) exam, as well as the North American Pharmacist Licensure Examination (NAPLEX) exam, which both establish a significant milestone in both the field of AI and medicine [12, 13].

In this study, we demonstrate the medical reasoning capabilities of these large language models by evaluating their performance on the North American Veterinary Licensing Examination (NAVLE), highlighting the potential for a transformative impact on veterinary medicine. The NAVLE, a standardized test administered by the International Council for Veterinary Assessment (ICVA), assesses the knowledge and skills of veterinary graduates seeking licensure to practice veterinary medicine in North America. Comprising 360 multiple-choice questions, the NAVLE covers topics such as animal health and welfare, diagnostic imaging, and pharmacology. Successful completion of the NAVLE is a prerequisite for licensure in the majority of North American states and provinces [14].

II. MATERIAL AND METHODS

The NAVLE exam sample employed in this study was obtained from the ICVA's website. This sample, provided by the National Board of Veterinary Medical Examiners, consists of 200 multiple-choice questions accompanied by an answer key [15]. Out of the total 200 multiple- choice questions, 36 questions included an image that was necessary to solve the answer, and 4 questions contained data tables to aid in finding the correct answer.

Three language models were utilized for this experiment: GPT-3, GPT-4, and Bard. The multiple-choice questions from the sample assessment were input into GPT-3 and GPT-4 using the ChatGPT user interface, while Bard was accessed via Google's user interface [16,17]. Even though numerous studies have shown that careful prompting such as using the chain of thought where the prompt guides the models to reason step by step can significantly increase the performance [18,19], for our work, since we aim to evaluate the inherent medical knowledge and reasoning abilities of the LLMs, we avoided using any advanced prompting techniques. Therefore, all models used in our work were only primed with the following prompt: "You are a veterinary school graduate taking the North American Veterinary Licensing Examination. For the following multiple-choice questions, indicate the correct choice."

In addition, it is worth noting that 36 questions in the sample assessment contained images that were integral to understanding the questions. However, since pure language models are incapable of directly interpreting images, we only considered the 164 text-only questions.

The responses generated by the language models were compared to the provided answer key. Each model was assigned a score out of 164 based on the number of correct responses. The results were then analyzed to evaluate the performance of the respective language models in the context of veterinary medical assessment.

III. RESULTS

From the assessment results, GPT-3 and Bard answered 63.4% and 61% of the questions correctly, respectively. GPT-4, a more advanced language model, demonstrated superior performance, correctly answering 89% of the questions (Table 1).

TABLE I
TEST RESULTS OF DIFFERENT LARGE LANGUAGE MODELS ON THE SAMPLE ASSESSMENT

Model	Raw Score	Percentage
Bard	100/164	61%
GPT-3	104/164	63.4%
GPT-4	146/164	89%

During the evaluation of LLMs using the sample NAVLE examination, Google's Bard ran into some difficulties responding to 4 out of the 164 questions. When entering these questions into Bard's interface, the only response it returned was "I am not programmed to assist with that". GPT-3 and GPT-4 were able to attempt all 164 questions.

IV. DISCUSSION

Our evaluation of the LLMs, including GPT-3, GPT-4, and Bard, demonstrates their potential in answering veterinary-related questions. Although the NAVLE exam utilizes a scoring method that accounts raw score with a slight adjustment depending on the question difficulty, it is reasonable to assume that an accuracy of greater than 70% would lead to a passing score [14,20]. Based on our evaluations, only GPT-4 achieved an accuracy of 70% or higher in answering the sample NAVLE exam questions.

As language models continue to evolve, it is expected that their performance on exams like NAVLE will improve [12, 13], as evidenced by GPT-4's superior performance compared to Bard and GPT-3, and the latest model that Bard uses outperformed its predecessor. Currently, there is no major conceptual obstacle in extending LLMs to be able to deal with multi-modal data. In particular, deep learning methods capable of analyzing and interpreting biomedical images and videos have been developed [1, 4].

The success of language models like GPT-4 in passing the NAVLE exam presents both opportunities and challenges. First, it must be noticed that passing the NAVLE is only one of the steps required to obtain veterinary licensure in the United States. The steps include completion of the DVM degree (or equivalent education) and a passing score on the NAVLE (North American Veterinary Licensing Examination). In addition, some states have additional requirements such as additional clinical competency tests and/or state jurisprudence exams.

On one hand, the use of language models could enhance diagnostic accuracy and improve treatment times for animals, leading to increased efficiency and economy of scales in veterinary practices. By harnessing the knowledge and expertise of these models trained on large amounts of data, far exceeding what any human individual can absorb, veterinarians could make more informed decisions and ultimately provide better care for their patients. The future of the veterinary field appears promising as AI technologies continue to advance. Language models can contribute to more accurate diagnoses and treatment plans, improved client communication, and enhanced efficiency within veterinary practices. Furthermore, AI may help minimize errors and bolster patient outcomes.

Additionally, language models may serve as valuable educational tools for veterinary students, enabling them to refine their diagnostic skills [21]. It is not surprising that these language models can successfully pass standardized medical tests which continue to be largely based on memorization and recall of published scientific data. Our study suggests that large language models such as GPT and Bard may potentially assist human learners in the veterinary medical education setting, as a prelude to future integration into medical decision-making where the ability to apply this knowledge in the clinical setting will be key.

Conversely, potential negative consequences warrant consideration. Large language models pose challenges in education, and how student's knowledge is to be assessed. More importantly, the incorporation of language models in the veterinary field might reduce the demand for veterinary professionals as some tasks become automated or outsourced. This could lead to job loss and diminished quality of care for animals. Moreover, ethical concerns surrounding machine-generated medical decisions, as well as potential errors or biases in the algorithms of these models, must be addressed. Ensuring that language models are responsibly and ethically integrated into veterinary medicine is crucial [22, 23].

Veterinary educators are moving towards implementing a competency-based veterinary education framework nationally [24]. Competency-based veterinary education is an approach modeled after competency-based medical education and is designed to prepare graduates for professional careers by confirming their ability to meet the needs of animals and the expectations of society. This approach focuses on outcomes-based and learner-centered education and assessment. A barrier to creating such new content is the human effort required to craft realistic clinical scenarios that explore complex medical concepts and emphasize critical thinking, rather than emphasizing the selection of the correct multiple-choice answer. As demand for this type of examination content continues to increase, generative language AI may offload this human effort by assisting in the question-explanation writing process or, in some cases, writing entire items autonomously [11].

In sum, on the positive side, the applications of AI in veterinary medicine will also create new jobs, for instance, AI software engineers, University faculty to develop new curriculum, State regulatory agencies to design and enforce new regulations, and develop and deploy new approaches to compliance, including automated testing methods. However, it is vital to address ethical concerns related to data privacy, bias, and job displacement. The veterinary community must collaborate with AI experts to employ this technology responsibly and effectively in the years to come [3].

REFERENCES

[1] Baldi P. (2021). Deep Learning in Science. Cambridge: Cambridge University Press. doi:10.1017/9781108955652232

[2] Dicks MR. A short history of veterinary workforce analyses. J Am Vet Med Assoc. 2013;242(8):1051-60. doi: 10.2460/javma.242.8.1051\

[3] Appleby RB, Basran PS. Artificial Intelligence in veterinary medicine. J AmVet Med Assoc. 2022; 260(8):819-824. doi: 10.2460/javma.22.03.0093.

[4] Ott J, Bruyette DS, Arbuckle C, Balsz D, Hecth S, Shubitz L, Baldi P. Detecting Pulmonary Coccidioidomycosis with Deep Convolutional Neural Networks. Machine Learning with Applications. 2021; 5,100040. doi:10.48550/arXiv.212.00280.

[5] Khan B, Fatima H, Qureshi A, et al. Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. Biomed Mater Devices. 2023; 8:1-8. doi:10.1007/s44174-023-00063-

[6] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. 2020.arxiv.org/abs/2005.1416

[7] OpenAI GPT-4 Technical Report. 2023. https://cdn.openai.com/papers/gpt-4.pdf.

[8] Thoppilan R, De Freites D, Hall J, et al. LaMDA: Language Models for Dialog Applications. 2022. https://arxiv.org/abs/2201.08239

[9] Vaswani A, Shazeer N, Parmer N, et al. Attention Is All You Need. 2017. https://arxiv.org/abs/1706.03762249

[10] Google ai palm 2. Google AI Available at: https://ai.google/discover/palm2.

[11] Kung TH, Cheatham M, Medenilla A, et al. Performance of Chat-GPT on USMLE: Potential for AI-assisted medical education using large language models. 2023. PLOS Digital Health 2(2): e0000198. https://doi.org/10.1371/journal.pdig.0000198

[12] Angel MC, Rinehart JB, Canneson MP, Baldi P. Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: A comparative study on the ABA exam. Published online 2023. doi:10.1101/2023.05.10.23289805

[13] Angel, M., Patel, A., Alachkar, A. & Baldi, P. Clinical knowledge and reasoning abilities of AI large language models in Pharmacy: A Comparative Study on the NAPLEX exam (2023). doi:10.1101/2023.06.07.544055

[14] NAVLE. ICVA Available at: https://www.icva.net/navle/.

[15] National Board of Veterinary Medical Examiners – self-assessment services:home https://csas.nbme.org/navlesa/Home.do.

[16] Open AI, ChatGPT Plus. 2023. https://openai.com/blog/chatgpt-plus

[17] Google Bard. 2023. https://bard.google.com/

[18] Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (2022), Arxiv. https://doi.org/10.48550/arXiv.2201.11903

[19] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners (2022), Arxiv. https://doi.org/10.48550/arXiv.2205.11916

[20] Feinberg RA, Jurich D, Lord J, Case H, Hawley J. Examining the Validity of the North American Veterinary Licensing Examination (NAVLE) Time Constraints. J Vet Med Educ. 2018. 45(3):381-387. doi: 10.3138/jvme.0217-026r.

[21] Hazarika I. Artificial Intelligence: Opportunities and implications for the Health Workforce. International Health. 2020. 12:241–245.

[22] Dahlin E. Are Robots Stealing Our Jobs? Socius: Sociological Research for a Dynamic World. 2019. doi: d0o.i.1o1rg7/71/02.1317870/2337181029381416928469

[23] Ogeer J. AI in veterinary medicine: From finding disease to predicting it. Veterinary Practice News. 2020. Available at: https://www.veterinarypracticenews.com/ai-diagnostics-december-2020/.

[24] Competency-Based Veterinary Education (CBVE). Available at: https://www.aavmc.org/programs/